

# Mex # 5 Fasttext Embedding

For the Mex # 5, I implemented a skip-gram model for fasttext with CUDA implementation to train 750 documents. The implementation was based off from *Enriching Word Vectors with Subword Information* from Bojanowski, P., et. Al. The model collects positive and negative word-context pairs from a sentence and computes the dot products of word vectors and context vectors, incorporating n-gram representations to capture sub-word information. Using stochastic gradient descent (SGD), the word vectors and context vectors are updated in parallel on the GPU for both positive and negative samples by launching CUDA kernels. The updated vectors are then transferred back from GPU to CPU where the vocabularies are updated separately.

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c$$
$$L = l(s(w, c)) + \sum_{n \in \mathcal{N}_{t,c}} l(-s(w, n_i))$$
$$\frac{\delta L}{\delta w} = l(s(w, c)) + \sum_{n \in \mathcal{N}_{t,c}} l(-s(w, n_i))$$

Where:

$l: x \mapsto \log(1 + e^{-x})$  is the logistic function

$\mathcal{N}_{t,c}$  is a set of negative samples

$G_w \subset \{1, \dots, G\}$  the set of n-grams appearing on  $w$ .

For positive samples:

$$w \leftarrow w - \eta(l(s(w, c)) - 1)c$$
$$c \leftarrow c - \eta(l(s(w, c)) - 1)w$$

For negative samples:

$$w \leftarrow w - \eta(l(s(w, n_i)) - 1)c$$

Where:

$w$  is the word vector

$c$  is the context vector

$\eta$  is the gradient

Results:

My custom model took 427 minutes and 19 seconds to train 1,035,580 sentences while the Gensim model only took 3 minutes and 2 seconds. I saw from my task manager, that around 50-60% of my GPU device (0) is used for my custom model while for the Gensim model, 100% of my CPU memory is used. What I'm getting at is maybe I haven't utilized all the threads that I have in the GPU unlike in Gensim where it was able to utilize most of my CPU memory. I was able to print the word, context and n gram vector embeddings on a .txt file as well as the comparison of the Gensim and my custom model on sample sentences. I was able to get a mean difference in the range of 0.2 to 0.3.

```
Epoch: 5/5, Sentence: 207102/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207102/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207103/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207104/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207105/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207106/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207107/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207108/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207109/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207110/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207111/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207111/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207112/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207113/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207114/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207115/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207115/207116, Remaining: 0 mins 0 secs: 100%|
Epoch: 5/5, Sentence: 207116/207116, Remaining: 0 mins 0 secs: 100%|
███| 1035580/1035580 [7:07:19<00:00, 29.76it/s]
-----DONE TRAINING-----
Training for 1035580 sentences completed in 427 minutes and 19 seconds.
Caching vectors...
```

## 1.) Word vector

```

-----Context Vectors-----
0 aab 0.78532544, 0.0564899, -0.01152891, 0.03673532, -0.083176, 0.07438426, -0.010247618, -0.08189934, 0.04596267, 0.035920236, -0.004301346, 0.05897605, 0.008467616, -0.04822142, 0.075
1 bat 0.0476424, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000, 0.0000000
2 bat 0.04043141, 0.07383347, 0.07202356, 0.043439428, -0.09022465, 0.0639355, 0.056563247, -0.037188665, -0.01564466, -0.07903279, 0.07642555, 0.05193225, 0.03222184, 0.04244454, -0.0372680
3 eat 0.05473747, 0.08194085, -0.00490085, 0.006220858, -0.0589266, 0.07034895, -0.0050978, -0.04144244, 0.02904214, 0.02695364, -0.037709151, 0.01841813, 0.0001875023, -0.0341
4 eat 0.02713209, 0.0025475, 0.05242312, -0.00521899, -0.00195026, -0.0575849, 0.04581474, 0.04582596, -0.07338892, 0.0011742818, 0.03544493, -0.01158512, -0.00476459, 0.04297459, 0.0
5 rac 0.0806931, -0.0372673, -0.09547947, 0.0871886, 0.09363225, -0.03169555, -0.09022036, -0.0981726, -0.05947167, -0.01132025, 0.05427695, -0.03720025, -0.04291343, -0.00582062, 0.06103
6 eat 0.09272466, 0.0925511, 0.00129439, -0.05270739, -0.0954764, -0.040992781, -0.07322505, 0.03994167, -0.0746025, 0.04095323, -0.06021446, 0.07482715, 0.04269946, -0.03527996, -0.066495
7 eat 0.0272751, 0.0025475, 0.05242312, -0.00521899, -0.00195026, -0.0575849, 0.04581474, 0.04582596, -0.07338892, 0.0011742818, 0.03544493, -0.01158512, -0.00476459, 0.04297459, 0.0
10 abstract 0.06890109, 0.04608343, 0.04229298, 0.04612523, -0.022122032, 0.02145007, -0.03214252, 0.0804078, -0.03429094, 0.04564606, -0.01138815, -0.08236525, -0.05076065, 0.0343219
12 eat 0.0743415, -0.04391734, 0.07149185, 0.06756102, -0.07633414, 0.0351172, 0.05321065, 0.03404453, -0.0253048, -0.0919416, 0.0487601, -0.05622805, -0.0434725, 0.07239551, -0.06131238
13 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
14 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
15 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
16 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
17 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
18 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
19 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
20 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
21 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
22 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
23 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
24 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
25 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 0.07855859, 0.06599497, -0.021726, -0.0005555, 0.05021255, 0.03776091, 0.01548700, -0.07950423, -0.07990466, 0.0
26 [em] 0.04677246, -0.01531663, -0.04081676, 0.09721246, 0.07121608, 
```

## 2.) Context vector

```
File Edit View Search Encoding Language Settings Tools Macro Run Plugins Window ?
File Edit View Search Encoding Language Settings Tools Macro Run Plugins Window ?

1 Word Vectors
2 cab 0.0448051, 0.0813024, 0.02944658, -0.0050408, 0.0859637, -0.01406238, 0.05763485, 0.00549595, 0.04057803, 0.01680902, 0.02535322, -0.09201231, -0.03104184, 0.084173635, -0.087665
3 0.043762304, 0.07667404, 0.09240057, 0.0643641, 0.0723081, 0.04060507, 0.02738464, 0.07428425, 0.02902985, 0.08924851, 0.04010413, 0.01794564, 0.08144224, 0.010919435, 0.026556252, -0.0726
4 -0.04343175, -0.03783347, 0.07202356, 0.046439428, -0.09023465, -0.03628557, -0.007188665, -0.01564646, -0.07903279, 0.07645525, 0.061953235, 0.023226184, 0.04244454, -0.03726
5 0.04343175, -0.03783347, 0.07202356, 0.046439428, -0.09023465, -0.03628557, -0.007188665, -0.01564646, -0.07903279, 0.07645525, 0.061953235, 0.023226184, 0.04244454, -0.03726
6 tra -0.052713208, -0.04995479, -0.043357853, -0.061879057, -0.03565306, 0.07843642, -0.02582454, 0.04382286, -0.0767382, 0.01742819, 0.035561483, -0.01505812, -0.031344313, -0.02647659, 0.0
7 tra 0.0808951, -0.0372873, -0.04087340, 0.0871488, 0.03983325, -0.02169955, -0.09022036, -0.0981726, -0.059947167, -0.03132052, 0.04572055, -0.02370205, -0.04293163, -0.04398282, 0.06103
8 tra 0.0808951, -0.0372873, -0.04087340, 0.0871488, 0.03983325, -0.02169955, -0.09022036, -0.0981726, -0.059947167, -0.03132052, 0.04572055, -0.02370205, -0.04293163, -0.04398282, 0.06103
9 tra -0.07712762, -0.02849189, -0.07564472, -0.076941, -0.08131466, -0.003102916, -0.08831806, -0.09669324, 0.02162828, -0.01384715, -0.015749458, 0.05360117, -0.0776396, 0.0872543, 0.05393
10 tra -0.07712762, -0.02849189, -0.07564472, -0.076941, -0.08131466, -0.003102916, -0.08831806, -0.09669324, 0.02162828, -0.01384715, -0.015749458, 0.05360117, -0.0776396, 0.0872543, 0.05393
11 abstract -0.065879434, 0.046081956, 0.04282914, -0.06410584, -0.002102164, -0.022141892, -0.032146923, 0.08803569, -0.0349012, -0.04562329, -0.0113847, -0.06823154, -0.05676234, 0.06493
12 abstract -0.065879434, 0.046081956, 0.04282914, -0.06410584, -0.002102164, -0.022141892, -0.032146923, 0.08803569, -0.0349012, -0.04562329, -0.0113847, -0.06823154, -0.05676234, 0.06493
13 [E] -0.043423995, -0.03685394, -0.03890383, 0.0866837, -0.04075266, -0.01819229, -0.02814957, 0.0974509, 0.06598592, -0.0773862, -0.06182839, 0.03190134, -0.0631056, -0.06060175, 0.0847871
14 [E] -0.043423995, -0.03685394, -0.03890383, 0.0866837, -0.04075266, -0.01819229, -0.02814957, 0.0974509, 0.06598592, -0.0773862, -0.06182839, 0.03190134, -0.0631056, -0.06060175, 0.0847871
15 [EN] 0.06930690, 0.0726555, 0.04092382, -0.07666772, 0.09541483, 0.004248345, -0.02687787, 0.071264975, 0.08915699, -0.04154203, 0.025712703, 0.08755151, -0.05897474, -0.02540186, -0.062083
16 [EN] -0.078506102, 0.09465579, -0.04270367, 0.09142674, -0.044755376, -0.04742785, -0.0871474, -0.018651037, 0.012631634, -0.0052119595, -0.03909093, 0.047602692, -0.019583033, 0.03910284, -0.06
17 [EN] -0.078506102, 0.09465579, -0.04270367, 0.09142674, -0.044755376, -0.04742785, -0.0871474, -0.018651037, 0.012631634, -0.0052119595, -0.03909093, 0.047602692, -0.019583033, 0.03910284, -0.06
18 [EN] 0.0049138716, 0.03777343, -0.00765447, 0.04830859, 0.04396955, -0.0462568, -0.061948252, -0.0751324, -0.07999526, -0.0858521, -0.0333339, -0.0282401, -0.0584789, -0.03021962, 0.0722
19 [EN] 0.0049138716, 0.03777343, -0.00765447, 0.04830859, 0.04396955, -0.0462568, -0.061948252, -0.0751324, -0.07999526, -0.0858521, -0.0333339, -0.0282401, -0.0584789, -0.03021962, 0.0722
20 [E] -0.09375936, -0.0851367, 0.0850463, 0.07321596, -0.0639179, 0.052849725, 0.09269266, -0.0712893, 0.031072818, -0.054140892, 0.06672625, 0.07454148, -0.04450892, 0.02695053, -0.03295993
21 [E] -0.09375936, -0.0851367, 0.0850463, 0.07321596, -0.0639179, 0.052849725, 0.09269266, -0.0712893, 0.031072818, -0.054140892, 0.06672625, 0.07454148, -0.04450892, 0.02695053, -0.03295993
22 lon -0.05796467, 0.0273773, 0.021009879, -0.005043174, -0.07379294, -0.060136948, 0.05933268, -0.02602116, -0.0697798, -0.04367262, 0.0414584, -0.03951043, -0.03951043, -0.0649553, 0.0871
23 lon -0.05796467, 0.0273773, 0.021009879, -0.005043174, -0.07379294, -0.060136948, 0.05933268, -0.02602116, -0.0697798, -0.04367262, 0.0414584, -0.03951043, -0.03951043, -0.0649553, 0.0871
24 lon -0.05796467, 0.0273773, 0.021009879, -0.005043174, -0.07379294, -0.060136948, 0.05933268, -0.02602116, -0.0697798, -0.04367262, 0.
```

### 3.) N gram vector

The image shows a Notepad++ editor window with a large block of assembly code. The title bar at the top reads "File Edit View Search Encoding Language Settings Tools Macro Run Plugins Window". The code is a mix of assembly instructions and comments, with a prominent "N-gran Vectors" section. The status bar at the bottom shows "Normal text file", "length: 36.89A bytes", "lines: 30403", "Ln:1 Col:1 Pos:1", "Windows (x64)", and "UTF-8". The code is a mix of assembly instructions and comments, with a prominent "N-gran Vectors" section. The status bar at the bottom shows "Normal text file", "length: 36.89A bytes", "lines: 30403", "Ln:1 Col:1 Pos:1", "Windows (x64)", and "UTF-8".

#### 4.) Comparison

[illegible]

